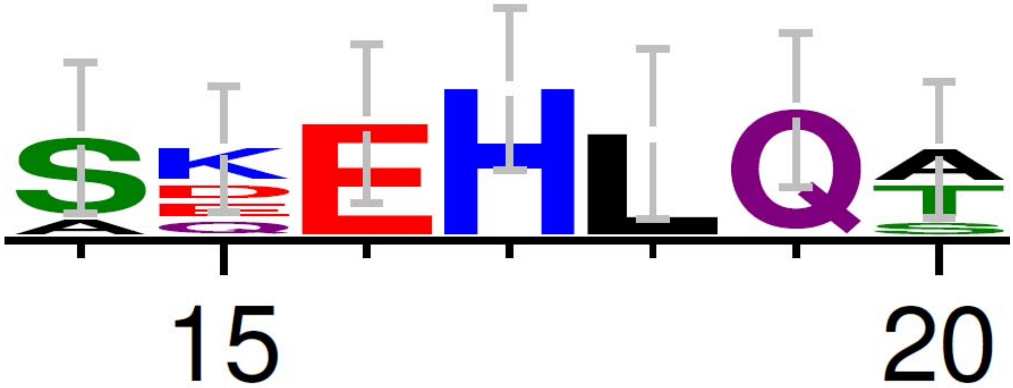
SLU

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Whole Genome Assembly, Annotation and Bioinformatics analysis of *Bacillus Amyloliquefaciens* strain *UCMB5113*”

Adnan Niazi



Department of Animal Breeding and Genetics
Examensarbete 335
Uppsala 2010

Master’s Thesis, 30 HEC

One-Year Master’s Programme in Biology – Bioinformatics

SLU

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Whole Genome Assembly, Annotation and Bioinformatics analysis of *Bacillus Amyloliquefaciens* strain *UCMB5113*"

Adnan Niazi

Supervisors:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Johan Meijer, SLU, Department of Plant Biology and Forest Genetics
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics

Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Cover picture: Adnan Niazi
Name of series: Examensarbete 335
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: Genome assembly, Bacillus amyloliquefaciens UCMB5113, plant growth and protection, bacterial biocontrol, antimicrobial genes

CONTENTS

Contents ... 1

1. ABSTRACT ... 2

2. INTRODUCTION ... 3

  2.1 Biocontrol ... 3

  2.2 Bacterial Biocontrol ... 3

  2.3 Bacillus and Biocontrol ... 4

    2.3.1 Plant Growth and Protection ... 4

    2.3.2 Induced Plant Resistance ... 5

  2.4 *Bacillus amyloliquefaciens* and Plant Resistance ... 6

3. AIMS ... 7

4. MATERIALS AND METHODS ... 8

  4.1 Isolation of the *UCMB5113* strain ... 8

  4.2 DNA Sequencing ... 8

  4.3 Genome Assembly and Annotation ... 10

    4.3.1 Genome Assembler ... 10

    4.3.2 Paired-end data preprocessing ... 10

    4.3.3 Mapping reads to the reference genome ... 11

    4.3.4 Base editing of draft genome ... 11

    4.3.5 De novo assembly of unaligned reads ... 11

    4.3.6 Gene prediction and gene annotation ... 12

5. RESULTS ... 13

  5.1 Mapping short reads data with the reference genome ... 13

  5.2 De novo assembly of unmapped reads ... 14

  5.3 Genes involved in biocontrol ... 15

6. DISCUSSION ... 19

7. CONCLUSION ... 20

8. ACKNOWLEDGMENT ... 21

9. REFERENCES ... 22

1. ABSTRACT

Bacillus amyloliquefaciens strain UCMB5113 is a Gram negative rhizo-bacterium, that produces antimicrobial compounds and induce plant basal defense with unknown genetic mechanism that suppress soil-borne plants and pathogens. Paired-end reads of length 75 bp with insert size of ~230 bp were sequenced from the genome of *UCMB5113* by Illumina multiplex technology. Reference-guided assembly of the paired-end reads was performed using *Bacillus amyloliquefaciens FZB42* as reference sequence which was found to be closely related to *UCMB5113* strain. A draft of *UCMB5113* genome obtained after mapping the reads with reference sequence containing 4,169 genes excluding 90 tRNA and 29 rRNA genes. The draft genome of *UCMB5113* harbors cluster of genes, which were found in *B. amyloliquefaciens FZB42* and involved in synthesis of secondary metabolites like surfactin, fengycin, bacillomycinD, bacillibactin, and bacilysin, by pathways not involving ribosomes. Initial analysis reveals that *UCMB5113* genome lacks an operon *NRS* involved in synthesis of putative secondary metabolite. The determination of draft genome sequence of *UCMB5113* provided very basic analysis of set of genes involved in plant bacterium interaction and plant growth and protection. Further analysis is required to understand the genetic mechanism of interaction and protection mediated by *UCMB5113*.

2. INTRODUCTION

Plants have to handle a changing environment with many different kinds of abiotic and biotic stress in order to survive. Factors like soil salinity, drought and inadequacy of nutrients limit the growth and productivity of plants and are referred to as abiotic stresses [1]. Biotic stress involves insect infestations and pathogen-mediated diseases which are deleterious for crop plants. According to an estimate, approximately 16% of the total crop yield is spoiled globally due to different diseases [2]. Over the past several years, different stress-preventing measures have been adopted which mainly include chemical fertilizers and pesticides, traditional breeding and more recently, genetic modification of plants. Such methods contributed significantly to the improvement of plant quality and productivity but they are costly tools for increasing the yield. In addition, side effects such as nutrient leakage, development of pesticide-resistant insect strains, and negative effects on the environment call for other means to maintain food safety and security. Consequently, researchers are developing alternative methods to overcome such problems. Bacteria-mediated biocontrol is one such strategy that has great promise to cope with such plant stresses and especially to withstand challenges imposed by climate change.

2.1 Biocontrol

Biocontrol is the use of other organisms to protect plants from diseases and pests. The protective organisms, like predatory insects, nematodes and bacteria can be introduced in a specific ecological niche to reduce the undesirable effects of plant damaging organisms. Use of such organisms to suppress the population of harmful pests and pathogens are referred to as Biological Control Agents (BCA). Biocontrol can be mediated in several ways for instance; releasing predatory insects in the field [3]; spraying bacteria and fungi on the aerial surface of plants that kills pests like *Bacillus thuringiensis* spores producing a toxin in the gut of insects [4]; and also by allowing bacteria to colonize the rhizosphere by seed treatment antagonizing attackers. Since a vast range of biocontrol methodologies exist the discussion will be limited to bacteria-mediated biocontrol.

2.2 Bacterial Biocontrol

Certain bacteria live in close association with other organisms such as plants; and form symbiotic, pathogenic, or neutral relationship with them. Symbionts such as endophytic bacteria and rhizobacteria help plants to acquire nutrients and provide protection, whereas pathogens utilize plant tissues and nutrients which affect growth and reproduction of the host plants. Rhizobacteria colonize the plant rhizosphere (the region of soil close to plant roots that is influenced by root secretions and associated soil microorganisms) and develop close physical and biochemical contacts with plants. Their primary source of nutrients is root exudates which mainly contain carbohydrates, organic acids and amino acids [5] and in return the bacteria help plants to tolerate stress [6]. This association between species is referred to as *mutualism*, in which both parties derive benefits. Rhizobacteria that support plant growth are termed as Plant-Growth-Promoting Rhizobacteria (PGPR) or more generally, Plant-Growth-Promoting Bacteria (PGPB). Many bacteria have also been found to improve plant defense to pathogens and insect pests and may or may not directly improve plant growth but are often also included in the PGPB category. The innate immune system of plants provides a basal defense activity by identifying the pathogen-associated molecular patterns (PAMPs) [7]. However, many plant pathogens have become capable, through co-evolution, to suppress such recognition or its downstream events and thereby cause infections. PGPB mediate plant growth and protection in different ways, for instance, making adequate amount of nutrients available (such as phosphate), by solubilization of iron (because iron may cause precipitation of soil phosphorous) [8]; by producing growth stimulating phytohormones (like auxins, gibberellins, and cytokinins) [9]; and by synthesizing different enzymes (like proteases and chitinases) that kill harmful microorganisms [10].

2.3 Bacillus and Biocontrol

Bacillus is a genus of gram positive and rod shaped bacteria, capable to produce endospores during stressful environmental conditions. Spores are very stable dormant structures that can survive for a long time even under harsh conditions. Among different bacterial species known to support plant growth and protection as shown in Table 1, *Bacillus* species like *B. cereus, B. subtilis, and B. amyloliquefaciens* are seemed to be most important.

Table 1. Some known biocontrol bacteria. (Dunn et al., 2003; Schisler et al., 2004; Rudrappa et al., 2008; van Loon et al., 1998).

Organism
Bacillus amyloliquefaciens
Bacillus subtilis
Bacillus polymoxa
Bacillus licheniformis
Bacillus cereus
Bacillus pumilis
Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas chlororaphis
Enterobacter agglomerans
Enterobacter cloacae
Serratia marcescens

2.3.1 Plant Growth and Protection

Members of the *Bacillus* genus produce different kind of growth inhibiting compounds to phytopathogens, such as kanosamine and zwittemycine A from *B. cereus* [11]. Similarly, strains of *B. subtilis* and *B. amyloliquefaciens* are also known to synthesize antimicrobial compounds, such as surfactin, iturin, and fengycin. These antibiotic compounds inhibit phytopathogen growth but are also necessary for root colonization, for instance, surfactin facilitates root colonization as well as controls the disease caused by *Pseudomonas syringae* on Arabidopsis plants [12, 13].

Another important way by which Bacilli provide protection to plants is through competition in the rhizosphere. They compete with soil pathogens for growth space and essential nutrients. Some of this effect is also achieved through synthesis of iron chelating compounds called siderophores, which allow synthesizing bacteria to scavenge and solubilize iron from the environment making it unavailable for deleterious bacteria [10]. Further, as mentioned above, the benefit of solubilization of iron is that it avoids precipitation of soil phosphorous which makes large amounts of phosphate available for plants.

2.3.2 Induced Plant Resistance

In addition to constitutive defense, plants have also evolved inducible resistance mechanisms for their defense to resist harmful bacteria and tolerate abiotic stress. Induced resistance to pathogens occurs in two forms, Systemic Acquired Resistance (SAR) and Induced Systemic Resistance (ISR), as shown in Figure 1. Certain bacteria produce salicylic acid (SA), a phytohormone, which makes plants tolerant against pathogens by inducing SAR. The SAR defense mechanism provide long-term systemic resistance to subsequent pathogen attack, and is dependent on SA which mediates activation of distal tissues containing pathogenesis related (PR) proteins that target different infectious organisms. Whereas, ISR signaling requires functional jasmonic acid (JA) and ethylene (Et), and does not depend on SA. ISR is induced in plants by certain strains of Bacillus and Pseudomonas bacteria only after infection by the pathogen. The enhanced ability of stimulating stronger resistance responses associated with ISR is called 'priming' [14].

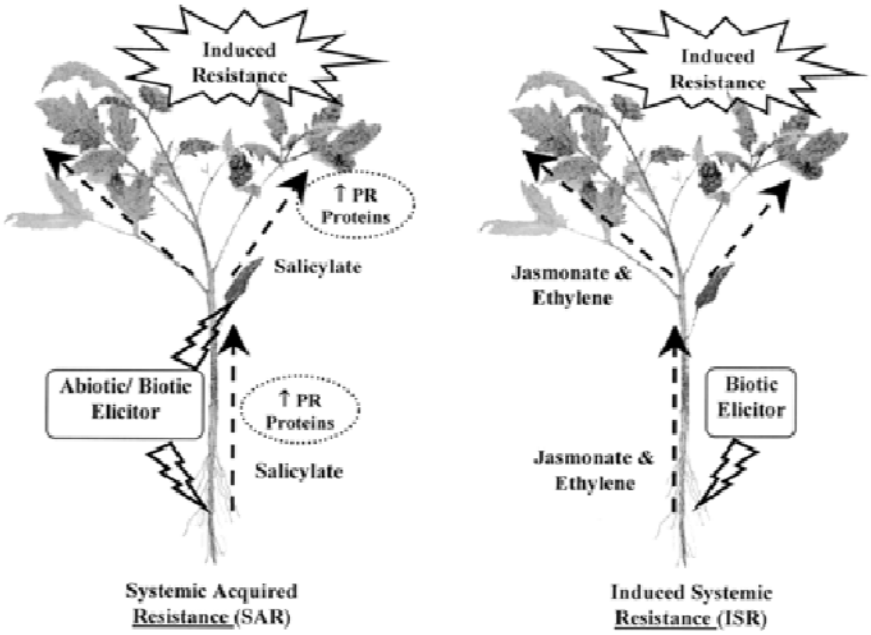


Figure 1. A schematic comparison of two forms of induced resistance in plants, both which lead to similar phenotypic responses. Both responses are intertwined molecularly, as demonstrated by their reliance on a functional version of the gene NPR1 in Arabidopsis thaliana. (Vallad GE and Goodman RM, 2004).

2.4 *Bacillus amyloliquefaciens* and Plant Resistance

Certain strains of *B. amyloliquefaciens* produce a wide range of antimicrobial compounds, for example, *B. amyloliquefaciens FZB42*. Genome sequence analysis of *B. amyloliquefaciens FZB42* revealed that substantial parts of the genome is responsible for non-ribosomal synthesis of secondary metabolites including lipopeptides and polyketides by some gene clusters with antimicrobial action [13,18]. Chen and collaborators reported that *B. amyloliquefaciens FZB42* produced three families of lipopeptides; surfactins, fengycins, and bacillomycins D, which are secondary metabolites known for antifungal activity [13]. These lipopeptides (LPs) have potential to interact with plant cells and induce plant resistance [12].

B. amyloliquefaciens FZB42 closely related strains, *UCMB 5036* and *UCMB 5113,* also produce some antifungal metabolites which unlike *FZB42* do not seem to play any role for the growth of plants but may provide protective ability on Brassica and Arabidopsis plants [16,17]. The main reason of choosing the *UCMB 5113* strain is because this strain seems to confer better protection to plants without producing too many effective antifungal compounds indicating it acts mainly through an ISR-based mechanism [17]. A great challenge is now to find and explain the genetic mechanism and biological pathways involved in the whole biocontrol process, from the establishment of bacteria in rhizosphere to increased plant vigour and stress tolerance.

3. AIMS

The overall aim of this project is to reveal and explain the genetic architecture that contributes to the ability of plants to overcome biotic and abiotic stress based on bacterial biocontrol. It is also necessary to do a comparative study of the genome of *B. amyloliquefaciens UCMB 5113* with other Bacillus species with known biocontrol ability and characterize their overall genomic differences. Furthermore, identification of potential genes involved in synthesis of metabolites supporting plant growth and protection provide a basis for future mechanistic studies of chemotaxis, colonization, priming and ISR. A good general understanding of molecular biology, programming skills and utilization of available bioinformatics resources are crucial components in this process where several tools have to be developed being in the research frontier.

4. MATERIALS AND METHODS

4.1 Isolation of the *UCMB5113* strain

The *B. amyloliquefaciens* strain *UCMB 5113* is a red pigmented bacteria isolated from soil of Karpaty mountains, Ukraine, which was identified as a member of the *B. amyloliquefaciens* group on the basis of phenotype [19]. *UCMB5113* strain, was chosen because it mediated pathogenic protection to oilseed rape in a recent study.

4.2 DNA Sequencing

Genomic DNA was extracted from *B. amyloliquefaciens UCMB5113*, and sequenced at the University Hospital, Molecular Medicine Laboratory, Uppsala University, through multiplexed sequencing process by Illumina technology. A total of 8,036,456 short insert paired-end reads of length 75bp were generated, with the average length of inserts at 230bp. The size of the paired-end library size ranged from 300 – 400bp including adapter and primer sequences as estimated from electropherogram (Figure 2a-2b) and ladder peak table (Table 2), and from standard curve plot (Figure 2c). All this information was provided by the sequencing company.

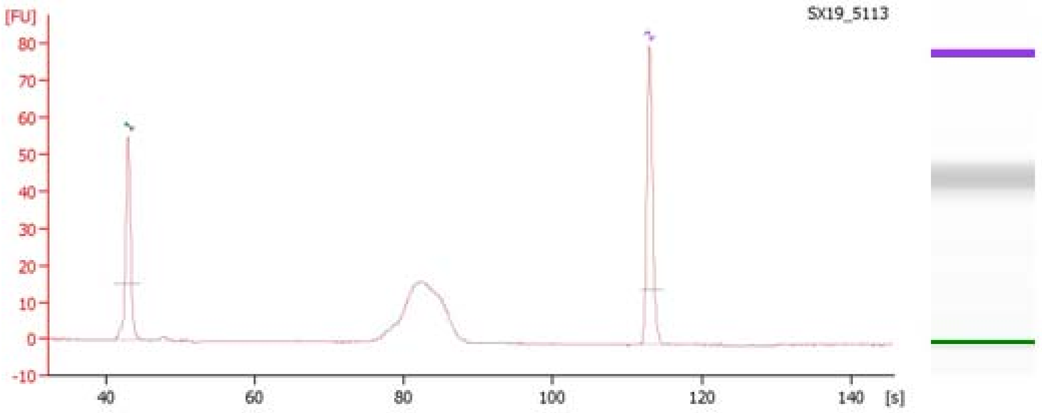


Figure 2a. Detection of sequence peak at 80-85sec elution time.

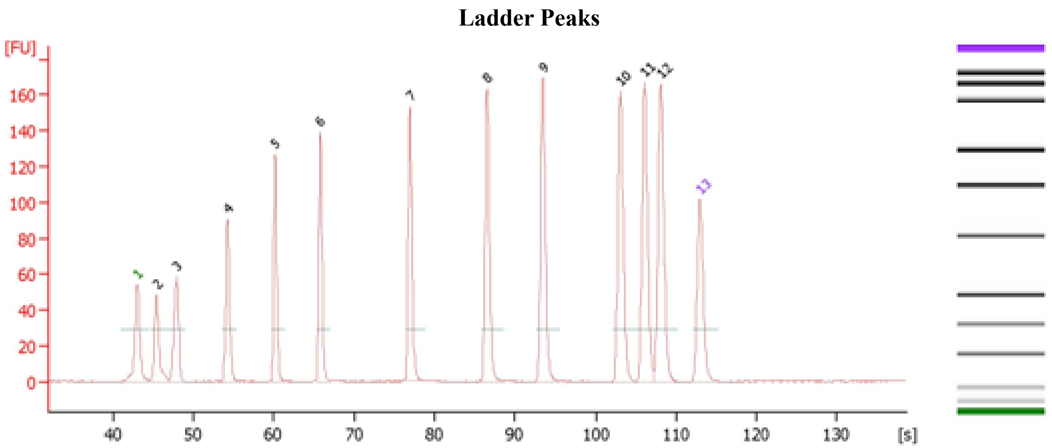


Figure 2b. Ladder peaks 7 and 8 near elution time 80-85secs corresponds to fragment length 300-400bp as described in Table 2.

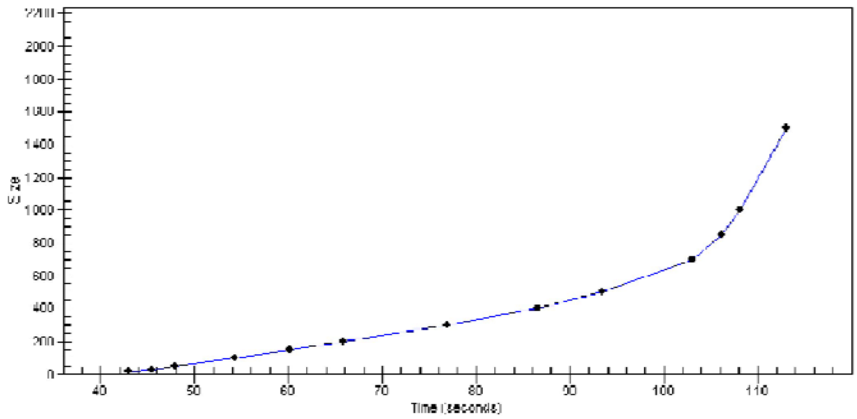


Figure 2c. Standard curve plot of sequencing illustrating reads fragment length is approximately 300-400bp at the interval of 80 sec.

Table2. Peak table for ladder indicating fragment length of paired-end reads.

Peak	**Size (bp)**	**Concentration (ng/µl)**	**Molarity (nmol/l)**	**Observations**
1	15	4.20	424.2	Lower Marker
2	25	4.00	242.4	Ladder Peak
3	50	4.00	121.2	Ladder Peak
4	100	4.00	60.6	Ladder Peak
5	150	4.00	40.4	Ladder Peak
6	200	4.00	30.3	Ladder Peak
7	300	4.00	20.2	Ladder Peak
8	400	4.00	15.2	Ladder Peak
9	500	4.00	12.1	Ladder Peak
10	700	4.00	8.7	Ladder Peak
11	850	4.00	7.1	Ladder Peak
12	1,000	4.00	6.1	Ladder Peak
13	1,500	2.10	2.1	Upper Marker

The reason to choose multiplexed sequencing technology by Illumina was due to its increased experimental throughput and reduced time and cost. Short insert paired-end libraries can be used for the detection of genetic variations like SNPs, indels, and genomic rearrangements. Paired-end reads also provide an advantage to identify repetitive sequence elements [20].

4.3 Genome Assembly and Annotation

4.3.1 Genome Assembler

There are several publicly available genome assemblers, such as Mosaik, AMOS, MIRA and MAQ. We chose *Mimicking Intelligent Reads Assembly* (MIRA version 3) for our project [21] because of its numerous features, such as; iterative assembly approach, handling repetitive elements, built-in clipping and masking of bad sequences like chimeric sequences and/or low quality sequences, tagging of regions of interest in the assembly, and various input and output formats.

The phases involved while assembling are as follows:

i. *Read comparison*: is the start of the assembly process. In this step, High Confidence Region (HCR) of each read is compared with HCF of every other read to find the possible overlaps among the reads which later on forms initial building graphs.

ii. *Systematic match inspection:* Building graphs that formed in previous step are reviewed with Smith-Waterman alignment algorithm [34]. Reads which do not fulfill the expected criteria are removed from initial building graph and accepted reads are inserted into alignment graphs.

iii. *Building contigs:* Contigs are built through alignment graphs which were formed in the systematic match inspection phase. Reads are added to the contigs based on their alignment with the consensus. Reads can be rejected during contig building if these introduce too many errors in the existing consensus.

iv. *Consensus sequence:* The consensus is formed after the above phases and written out to standard file formats for post assembly processes.

4.3.2 Paired-end data preprocessing

i) Format Conversion of Data

The Illumina reads that were obtained from the bacterial genome were paired-ends and came in two files, one for each end. They were initially in *scarf ASCII* format containing sequence identifiers, reads, and their quality values ranging from 64-104 ASCII characters. The format needed to be converted into FASTQ because most of the assemblers recognize FASTA/Q format as standard input. For this reason, we wrote a small script in Perl which could convert data from scarf to Illumina 1.3 FASTQ format.

ii) Contamination Screening

The primer and adapter sequences used for Illumina sequencing are listed in Table 3. If the read data would contain adapter sequences it could affect alignment of the reads. To avoid this, we performed screening of reads for adapter sequences using *ssaha2*, an alignment program from Sanger Center (with parameters “ssaha2 -output ssaha2 -kmer 8 -skip 1 -seeds 1 -score 12 -cmatch 9 -ckmer 6”).

Table 3. List of adapter sequences used during sequencing.

	Sequence (5´- 3´)
Adapter	GATCGGAAGAGCACACGTCT
Adapter	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

4.3.3 Mapping reads to the reference genome

Paired-end reads of *UCMB5113* after processing were mapped against the reference genome of *B. amyloliquefaciens FZB42* (available at NCBI) using the MIRA3 assembling software. The reference sequence in GenBank format (gbf) was used because the format has information regarding annotation which MIRA can make full use of after assembly. The assembler was run in mapping mode with the following parameters:

"--project=fz5113 --job=mapping,genome,accurate,solexa -AS:nop=2 -SB:ls=yes:bsn=FZB42: bft=gbf:lb=yes:bbq=30 SOLEXA_SETTINGS -CO:msr=no -GE:uti=no:tismin=180:tismax=280 -CL:msvs=yes: -LR:lsd:yes:ft=fastq:fqqo=64"

4.3.4 Base editing of draft genome

Consensus sequence containing IUPAC codes were edited by visualizing the assembly using an assembly viewer TABLET [22]. The IUPAC bases with the most convincing nucleotide bases were edited considering alignment coverage and quality nearby. The editing of IUPAC was done manually in consensus sequence *fasta* file.

4.3.5 De novo assembly of unaligned reads

De novo assembly was performed of the reads which we were unable to map with the reference genome because some of them were too short to be mapped after clipping bad quality bases, some appeared to be chimeric and/or contaminated reads, and some due to absence of sequence in the reference genome. Such reads were sent to a debris file by the assembler. The debris file only contained sequence identifiers of the unmapped reads, unaligned reads for which we need to extract debris from original set of reads. In order to extract unmapped reads, we loaded debris information and dataset in MySQL database. The data set was converted from *scarf ASCII* to comma separated *scarf numeric* format using Perl script, because the database could not load non-numeric qualities into the table. The reason we made *Scarf numeric* format comma separated was that sequence identifier, reads, and qualities could be loaded into individual columns of the table. Debris reads were extracted and exported in scarf numeric format which we again converted to *ASCII* and finally to FASTQ format.

De novo assembly of debris reads was completed with the parameters: ("--project=fz5113 --job=denovo,genome,accurate,solexa SOLEXA_SETTINGS -GE:tismin=180:tismax=280"), where 'tismin' and 'tismax' is approximately 20% deviation from the insert size (230bp) of paired-end reads.

4.3.6 Gene prediction and gene annotation

The draft genome was provided to BASys (Bacterial Annotation System), a web-based annotating pipeline [23], to predict functional elements in the genome and attach biological information to them. Unlike other annotation systems, BASys provides genomic information in both textual and graphical form. The process was accomplished in three parts: (*i*) submission of genomic data through web interface; (*ii*) analysis of genome for annotation in annotation engine; and (*iii*) annotation report generation in textual and graphical format.

Draft was provided to the RNAmmer program [24] for the annotation of rRNA genes while tRNAs genes were annotated using the tRNAscan-SE software [25].

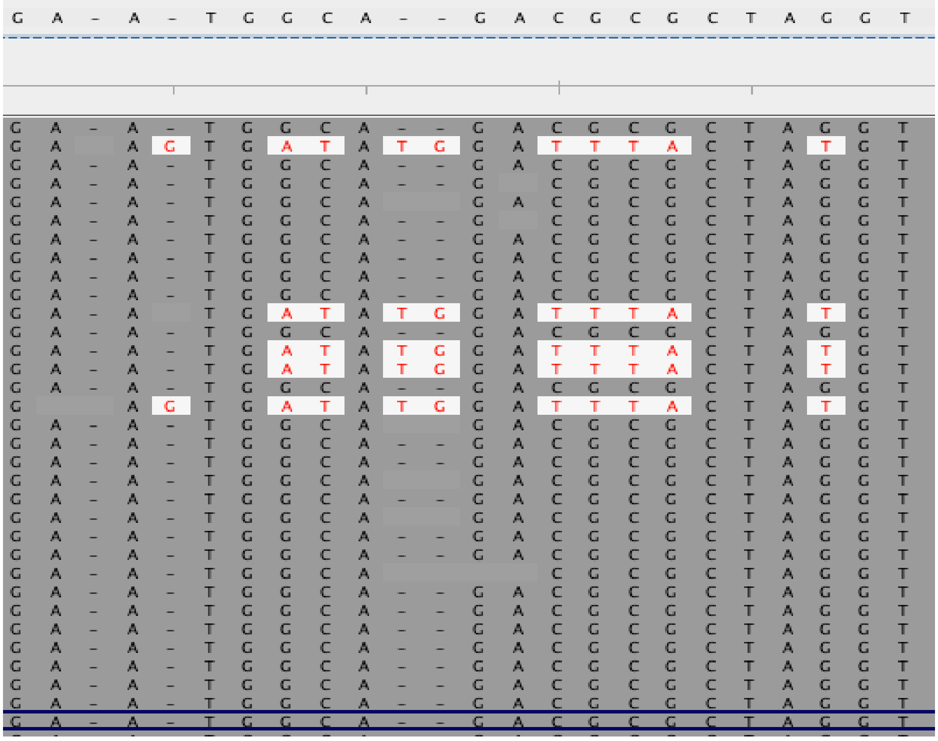
5. RESULTS

5.1 Mapping short reads data with the reference genome

The paired-end reads obtained from the Illumina genome Analyzer with only 0.23% observed adapter/primer contamination were mapped to the published *B. amyloliquefaciens* reference genome of *FZB42* (available at NCBI) using MIRA software. Libraries ranging from insert size 170-200bp are reliable in terms of adapter contamination [26], and showed no significant impact on the assembly of *UCMB5113*. Of the total reads, 93.38% reads could be mapped with an average sequencing coverage of 145-fold across the entire genome. It took ~16h (for 2 iterations) on 8 nodes each with 2 Quad/Core Intel 2.3 GHz processors, and 48GB memory. The consensus sequence or *draft genome* with average quality 34 was produced from reference guided assembly of short reads. The results are summarized in Table 4.

Table 4. Summary of reads assembled, coverage, and quality score.

Strain	Total reads (million)	Mapped reads (%)	Coverage (fold)	G+C (%)	Quality score 0-40
UCMB5113	8.36	93.38	145	46.49	34



```
G A - A - T G G C A - - G A C G C G C T A G G T

G A - A - T G G C A - - G A C G C G C T A G G T
G A   A G T G A T A T G G A T T T A C T A T G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G   C G C G C T A G G T
G A - A - T G G C A     G A C G C G C T A G G T
G A - A - T G G C A - - G   C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A   T G A T A T G G A T T T A C T A T G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G A T A T G G A T T T A C T A T G T
G A - A - T G A T A T G G A T T T A C T A T G T
G A - A - T G G C A - - G A C G C G C T A G G T
G     A G T G A T A T G G A T T T A C T A T G T
G A - A - T G G C A     G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A     G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A     G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A         C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
G A - A - T G G C A - - G A C G C G C T A G G T
```

Figure 4a. Illustrating strong repeat markers (SRMc) in the assembly.

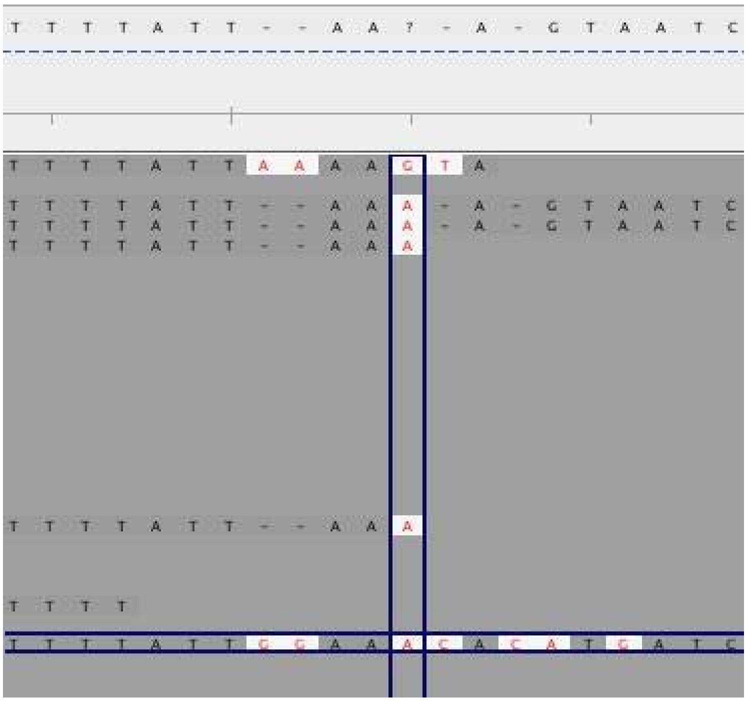

T T T T A T T - - A A ? - A - G T A A T C
T T T T A T T A A A A G T A
T T T T A T T - - A A A - A - G T A A T C
T T T T A T T - - A A A - A - G T A A T C
T T T T A T T - - A A A
T T T T A T T - - A A A
T T T T
T T T T A T T G G A A A C A C A T G A T C

Figure 4b. Base conflict due to low coverage and surrounding indels.

A consensus sequence of 3,918,894bp length was obtained from mapping assembly. Some regions in the consensus which needed attention were tagged. Of the total number of tags, 200 were regions of possible strong repeat markers and 132 were weak markers (SRMc and WRMc) as illustrated in Figure 4a. The regions with missing coverage or deletions were 524 and tagged as MCVc in the genome. The assembler at some stages was unsure about resolution of base conflicts during assembly due to repetitive elements in the genome or misassembling of reads, and IUPAC codes were inserted in the genome. There were 799 such regions which were tagged as IUPc, Figure 4b. SNPs were tagged as SROc in the consensus. Identified SNPs contained 38783 substitutions, and 1021 1-3bp indels in the genome.

5.2 De novo assembly of unmapped reads

De novo assembly of 4.98% unmapped reads with MIRA3 produced 3,205 contigs. The total size of produced contigs was 437,925bp and that of 85 contigs greater than 500bp in length was 174,873bp. The size of largest contig produced was 12,833bp. All statistics regarding produced contigs are shown in Table 3.

Table 3. Summary of contigs produced by MIRA3 with De novo assembly.

Debris reads (million)	Total Contigs	Total Contigs >500bp	Largest contig (bp)	Avg. Coverage (fold)
0.4	3205	85	12833	124.21

5.3 Genes involved in biocontrol

BASys provided automated, in-depth annotation of bacterial genome with more than 30 different tools and databases to produce approximately 60 separate annotations for each gene, map shown in Figure 5a,b. A total of 4,169 genes were predicted and annotated, excluding 29 rRNA and 90 tRNA genes, from the *UCMB5113* draft genome. Whereas, 3,693 genes were reported in the reference *B. amyloliquefaciens FZB42* genome [27].

Mapping of reads to the reference genome in the regions where non-ribosomal polypetide synthases (NRPS) and polyketide synthases (PKS) gene clusters were located showed existence of the same organization in *UCMB5113*, except for the gene cluster *Nrs*, synthesizing putative peptides, that was not present in *UCMB5113*. The NRPS and PKS gene clusters are involved in the synthesis of secondary metabolites in *FZB42* [27] listed in Table 5. An alkaline serine protease (AprE) that probably kills plant pathogenic nematodes was found in the *UCMB5113* genome and displayed high identity (99.7%) with *B. amyloliquefaciens FZB42* [21]. The genome of *UCMB5113* contains some genes (Table 6), known to be necessary for plant growth and plant-bacterium interaction. Moreover, the genes *YsnE* and *YhcX* (likely to be involved in synthesis of plant growth promoting IAA) showed high similarity with *FZB42*. Phylogenetic analysis of some genes showed that *UCMB5113* is closely related to *B. amyloliquefaciens FZB42* (Figure 6). BLASTX analysis of *de novo* produced contigs from unmapped reads could not bring any significant similarity with any protein coding sequences.

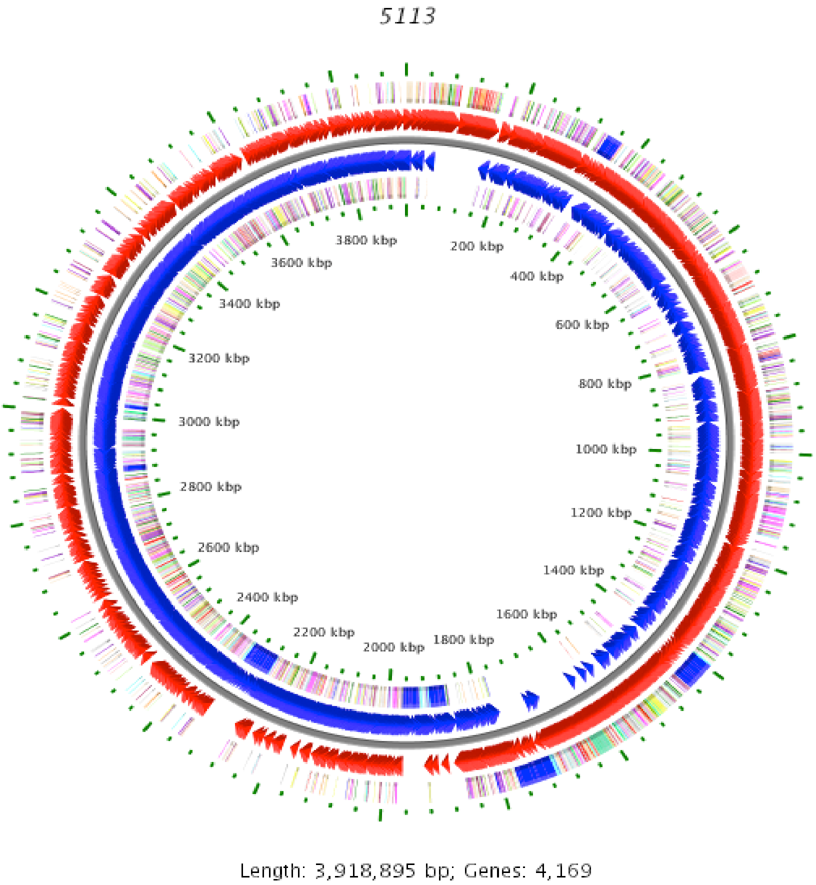


Figure 5a. Map view of genes encoding proteins and functional RNAs in the genome. Filled circular red and blue lines represent forward and reverse strands, respectively. Whereas, others demonstrating COG functional categories in the genome.

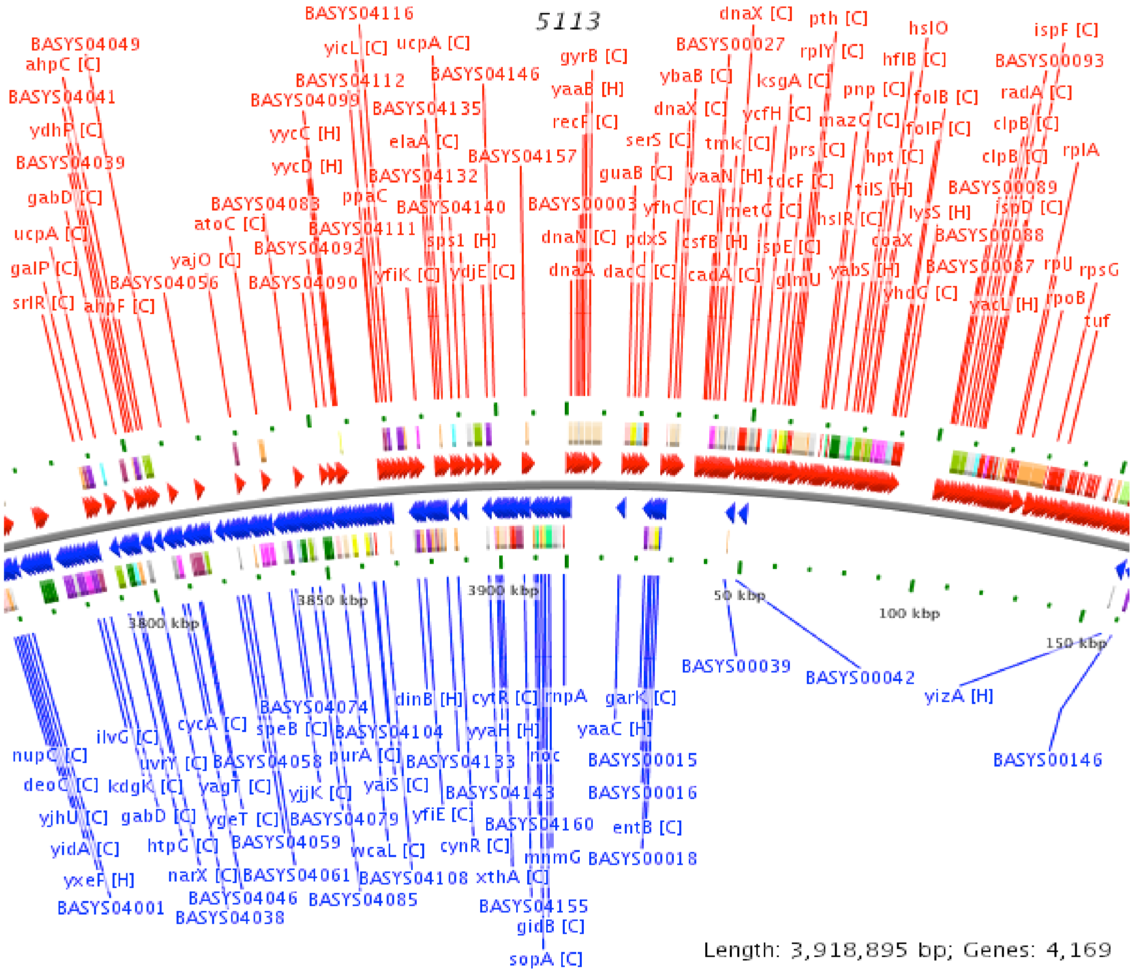


Figure 5b. Expanded view of annotation map illustrating genes predicted in UCMB5113 through the BASys annotation pipeline.

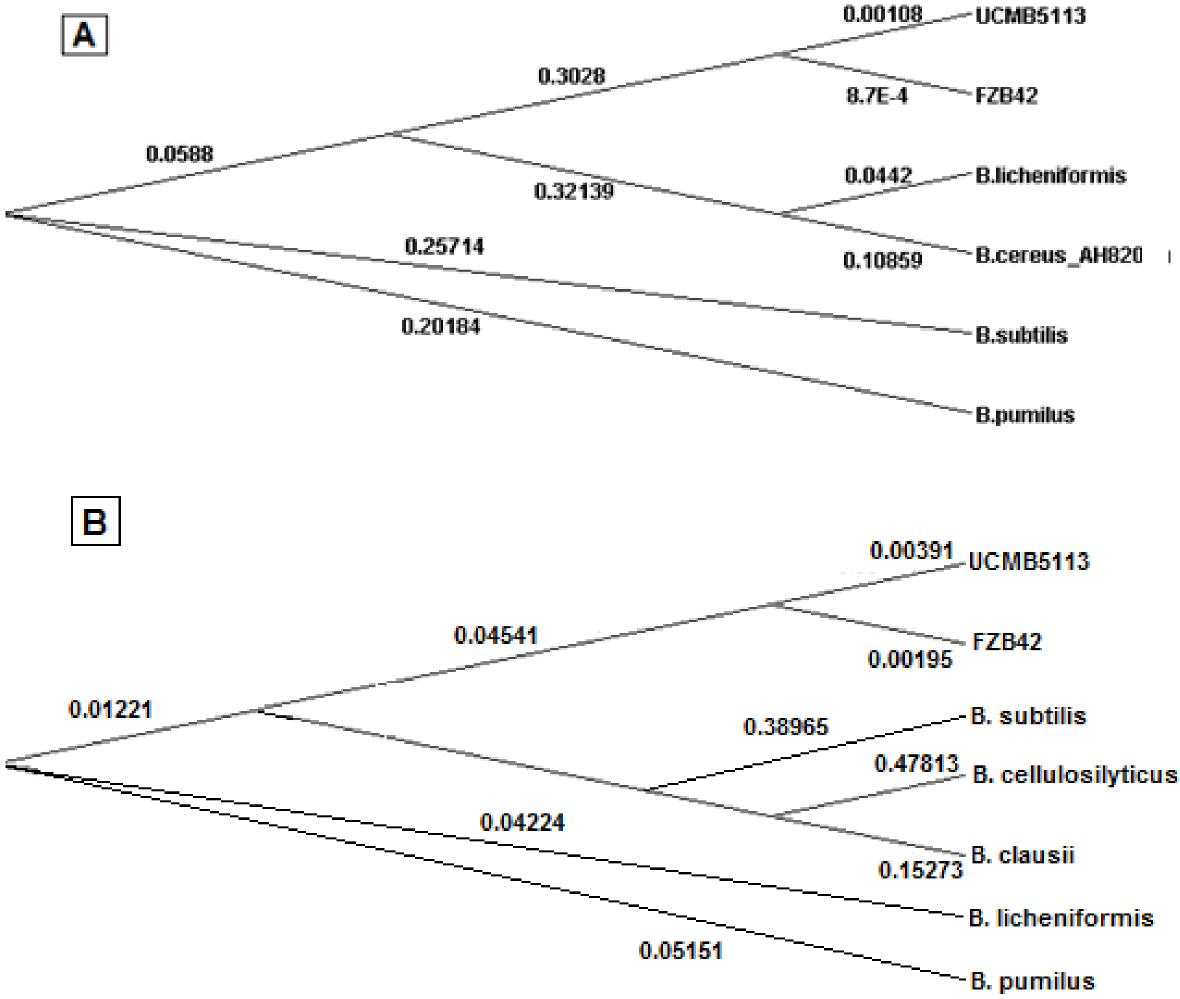


Figure 6. (a) Phylogenetic analysis of a 16S rRNA gene showed that UCMB5113 and B. amyloliquefaicens are closely related to each other. (b) Phylogram of the putative growth promoting YhcX gene showed high similarity with B. amyloliquefaciens FZB42.

Table 5. NRPS and PKS gene clusters involved in synthesis of secondary metabolites in B. amyloliquefaciens FZB42 and UCMB5113.

Compound	**Activity**	***UCMB5113 genes***	**Enzyme**	**Identity with *FZB42* (%)**
Surfactin	Antifungal	*srfABCD,ycxC,ycxD, sfp, yczE*	NRPS	99%
BacillomycinD	Antifungal	*bmyCBAD*	NRPS/PKS	98%
Fengycin	Antifungal	*fenABCDE*	NRPS	98%
Bacillibactin	Iron Siderophores	*dhbABCEF*	NRPS	98%
Bacilysin/anticapsin	Antibacterial	*bacABCDE, ywfG*	NRPS	99%
Macrolactin	Antibacterial	*mlnABCDEFGHI*	PKS	98%
Bacillaene	Antibacterial	*baeBCDE, acpK, baeGHIJLMNRS*	PKS/NRPS	98%
Difficidin	Antibacterial	*dfnAYXBCDEFGHIJKLM*	PKS	98%

Table 6. Genes probably involved in plant bacterium interaction and plant growth.

Gene	**Product**	**Identity with *FZB42* (%)**
swrA	Essential for swarming motility	90%
swrB	Essential for swarming motility	97%
Spo0A	Involved in initial stage of biofilm formation	98.9%
yfiQ	Probably involved in biofilm formation/surface adhesion	92.7%
ycdH	Probably involved in biofilm formation/surface adhesion	95%
ysnE	Probably involved in plant growth by synthesizing IAA	85.6%
yhcX	Probably involved in plant growth by synthesizing IAA	99%

6. DISCUSSION

Sequence analysis of 16S rRNA gene for phylogenetic studies shows that *UCMB5113* is closely related to *B. amyloliquefaciens FZB42*. We found that *UCMB5113* like *B. amyloliquefaciens FZB42* also contain gene clusters responsible for the production of polyketides such as bacillaene, difficidin, and macrolactin. Furthermore, gene clusters involved in production of lipopetides; surfactin, fengycin, and bacillomycin-D are also present in the genome. But how conserved they are relative to *FZB42* and how similar the corresponding antibiotics then can assumed to be demands additional analysis. The overestimated ORFs in *UCMB5113* as compared to *B. amyloliquefaciens FZB42* could be either due to presence of some genes which are not found in *B. amyloliquefaciens FZB42* or could be erroneous ORF prediction by the annotation pipeline or that the original *FZB42* genome sequence is not entirely correct in this region. According to our initial analysis, the gene *dhbD* in the bacillaene gene cluster has been mistakenly reported by Chen and collaborators [27]; it does not seem to exist in the genomes of *B. amyloliquefaciens FZB42*, *B. subtilis* 168, or in *UCMB5113* as well. The *UCMB5113* strain does not seem to be actively involved in plant growth, at least of oilseed rape, but the presence of growth promoting genes homologous to *B. amyloliquefaciens FZB42* suggests that either these genes are not induced when in contact with plants or the level of production is low.

Mapping assembly of reads is significant in determining polymorphisms and the genomic structure of an organism but it cannot assemble read sequences which are not present in the reference genome or having high degree of nucleotide sequence divergence. The draft genome of *UCMB5113* contains several large and small coverage missing regions (gaps) which need to be analyzed to ascertain the presence or absence of certain genes. Therefore, a combination of mapping and *de novo* assembly of the whole genome paired-end reads will be a good approach for finding missing or novel genes, resolving complex repetitive regions, and filling the gaps. Remaining gaps can also be covered with the help of PCR techniques.

Finalizing the consensus genome is also essential to solve errors in the process. The tags in the assembly are understandable by different finishing tools such as GAP4 from the Staden package [28], and CONSED [29], both of these allow manual editing. Unfortunately, we failed to inspect and finish the tags attached by MIRA in the assembly because of unsupported version of a tool *caf2gap* (32-bit) used for the conversion of *CAF* assembly file to *gap4* database file format required by GAP4. But since we are in the middle of the process such issues will be resolved in the future.

Completion of the *UCMB5113* genome will reveal its potential of producing secondary metabolites for developing agrobiotechnological agents. The genome should be equally valuable in explaining the complex interactions between PGPB bacteria and plants, *e.g.* the role of biofilm formation for bacterial adhesion to plant roots, as well as other biological processes.

7. CONCLUSION

The determination of a draft genome sequence of *UCMB5113* through mapping assembly provided information about a set of genes related to secondary metabolites. Annotation of the draft genome revealed that *UCMB5113* also harbours the operons involved in synthesis of antimicrobial compounds described in *B. amyloliquefaciens FZB42*. Further analysis is required to understand which genes made the strain *UCMB5113* capable enough to confer better plant protection without producing too many effective anti-fungal compounds.

8. ACKNOWLEDGMENT

First of all I am very thankful to Higher Education Commission (HEC), Pakistan for the complete financial support during this journey. I would also like to thank my supervisor Erik Bongcam-Rudloff for his support and guidance whenever I needed. Thanks to Johan Meijer for giving me clear view and for all interesting discussions related to my work.

Thank you all my friends, especially Dawood Rashid, Ali Asad, Usman Ahmed, Adil Gul, and Raheel Sardar. I appreciate your well wishes and moral support to me.

Lastly, special thanks to my parents and my siblings. It was all possible only due to them, because their prayers and well wishes were always with me.

9. REFERENCES

1. Kramer, P.J and Boyer, J.S (1995) Water relations of plants and soils. American Press

2. Neal Evansl, Michael H. Butterworth, Andreas Baierl, Mikhail A. Semenov, Jon S. West, Andrew Barnes, Dominic Moran and Bruce D. L. Fitt (2010). The impact of climate change on disease constraints on production of oilseed rape. Food Security. 2, 143-156

3. Wang B, Ferro D and Hosmer D (1999) Effectiveness of *Trichogramma ostriniae* and *T. nubilale* for controlling the european corn borer *Ostrinia nubilalis* in sweet corn. Entomol. Exp. Appl. 91, 297-303

4. Roh JY, Choi JY, Li MS and Jin BR, Je YH (2007) *Bacillus thuringiensis* as a specific, safe, and effective tool for insect pest control. Microbial Biotechnol. 17, 547-559

5. Lugtenberg B, Dekkers L and Bloemberg G (2001) Molecular determinants of rhizosphere colonization by Pseudomonas. Annu. Rev. Phytopathol. 39, 461-490

6. Yang J, Kloepper JW and Ryu CM (2009) Rhizosphere bacteria help plants tolerate abiotic stress. Trends Plant Sci. 14, 1-4.

7. He P, Shan L and Sheen J (2007) Elicitation and suppression of microbe associated molecular pattern-triggered immunity in plant-microbe interactions. Cell. Microbiol. 9, 1385-1396.

8. Gyaneshwar P. et al. (2002) Role of soil microorganisms in improving P nutritions of plants. Plant Soil 245, 83-93

9. Varma A, Abbot L, Werner D and Hampp R (eds) (2004) Plant surface microbiology, Springer, Berlin

10. Whipps J (2001) Microbial interaction and growth in the rhizosphere. J. Exp. Bot. 52, 487-511

11. Emmert EAB, and Handelsman J (1999) Biocontrol of plant disease: a (Gram-) positive perspective. FEMS Microbiol. Lett. 171, 1-9

12. Ongena M and Jacques P (2008) Bacillus lipopeptides: versatile weapons for plant disease biocontrol. Trends Microbiol. 16, 115-125

13. Chen XH et al. (2006) Structural and functional characterization of three polyketide synthase gene clusters in *Bacillus amyloliquefaciens* FZB 42. J. Bacteriol. 188, 4024-4036

14. Conrath U, Pieterse CMJ and Mauch-Mani B (2002) Priming in plant-pathogen interactions. Trends Plant Sci. 7, 210-216

15. Vallad GE and Goodman RM (2004) Systemic acquired resistance and induced systemic resistance in conventional agriculture. Crop Sci. 44, 1920-1934

16. Danielsson J, Reva O, and Meijer J (2007) Protection of oilseed rape (*Brassica napus*) toward fungal pathogens by strain of plant-associated *Bacillus amyloliquefaciens.* Microbiol Ecol. 54, 134-140

17. Danielsson J (2008) Bacillus based biocontrol on Brassica. Acta Universitatis agriculturae Sueciae nr 2008:40.

18. Koumoutsi A et al. (2004) Structural and functional charaterization of gene clusters directing nonribosomal synthesis of bioactive cylic lipopeptides in *Bacillus amyloliquefaciens FZB42*. J. Bacteriol. 186, 1084-1096

19. Reva ON, Dixellius C, Meijer J, Priest FG (2004) Taxonomic characterization and plant colonizing abilities of some bacteria related to *Bacillus amyloliquefaciens* and *Bacillus subtilis*. FEMS Microbiol. Ecol. 48, 249-259

20. Illumina-Sequencing Technology, http://www.illumina.com/

21. Chevreux et al. (1999) Genomic sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

22. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F and Marshall D (2010) Tablet—next generation sequence assembly visualization. Bioinformatics. 26, 401-402.

23. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, and Wishart DS (2005) BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 1;33(Web Server issue):W455-9.

24. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T and Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100-3108.

25. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955-964

26. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.*2008; **36**(19):e122. doi: 10.1093/nar/gkn502

27. Chen XH *et al.*(2007), Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens FZB42*. Nat Biotechnol.25, 1007-1014

28. Bonfield, J. K., Smith, K. F. and Staden, R. (1995), A new DNA sequence assembly program. Nucleic Acids Research, 23(24), 4992-4999.

29. Gordon D, Abajian C and Green P(1998) Consed: a graphical tool for sequence finishing. Genome Research 8, 195-202

30. Dunn AK, Klimowicz AK and Handelsman J (2003) Use of a promoter trap to identify *Bacillus cereus* genes regulated by tomato seed exudate and a Rhizosphere resident, *Pseudomonas aureofaciens*. Appl. Environ. Microbiol. 69, 1197-1205.

31. Schisler D, Slininger P, Behle R and Jackson M (2004) Formulation of Bacillus spp. for biological control of plant diseases. Phytopathology 94, 1267-1271.

32. Rudrappa T, Biedrzycki M, Bais HP (2008) Causes and consequences of plant-associated biofilms. FEMS Microbiol. Ecol. 64, 153-166.

33. Van Loon LC, Bakker PAHM, Pieterse CMJ (1998). Systemic resistance induced by rhizosphere bacteria. Annu. Rev. Phytopathol. 36, 453-483.

34. Smith, Temple F.; and Waterman, Michael S. (1981) Identification of Common Molecular Subsequences. Journal of Molecular Biology 147: 195–197.